Incorporating Constraints into Matrix Factorization for Clothes Package Recommendation

Conference or Workshop Item

Incorporating Constraints into Matrix Factorization for Clothes Package Recommendation

Agung Toto Wibowo*
Computing Science / Informatics Engineering
University of Aberdeen / Telkom University
wibowo.agung@abdn.ac.uk /
agungtoto@telkomuniversity.ac.id

Advaith Siddharthan
Knowledge Media Institute
The Open University
advaith.siddharthan@open.ac.uk

Judith Masthoff
Computing Science
University of Aberdeen
j.masthoff@abdn.ac.uk

Chenghua Lin
Computing Science
University of Aberdeen
chenghua.lin@abdn.ac.uk

ABSTRACT

Recommender systems have been widely applied in the literature to suggest individual items to users. In this paper, we consider the harder problem of package recommendation, where items are recommended together as a package. We focus on the clothing domain, where a package recommendation involves a combination of a "top" (e.g. a shirt) and a "bottom" (e.g. a pair of trousers). The novelty in this work is that we combined matrix factorisation methods for collaborative filtering with hand-crafted and learnt fashion constraints on combining item features such as colour, formality and patterns. Finally, to better understand where the algorithms are underperforming, we conducted focus groups, which lead to deeper insights into how to use constraints to improve package recommendation in this domain.

CCS CONCEPTS

•**Information systems → Recommender systems;**

KEYWORDS

Constraints, Package Recommendation, Matrix Factorization, Clothes Domain

1 INTRODUCTION

Research in recommender systems (RS) has been influenced by e-commerce websites (e.g., Amazon and Netflix) that produce recommendations for their users by exploiting implicit and explicit user interaction data from their systems [9]. For instance, implicit feedback may be gleaned from browsing or buying behaviors of a user, whereas explicit feedback might be gathered each time a user provides a rating for or comments about an item. These interactions together with users' personal data and items' descriptions are valuable input for recommender system approaches such as collaborative filtering, [4, 14], content based filtering [9], and hybrid methods [16].

Most research on generating recommendations focus on predicting ratings by a user for individual items. However, there are many cases where recommendations as a package better serve users' need. For example, a collection of music tracks as a play list [2], a collection of plants to support a particular animal species [17], a set of travel destinations as a tour package [5, 8], or in the clothing domain, a combination of a top (e.g. a shirt) and a bottom (e.g. trousers) [18].

One approach to package recommendations involves optimization. For example, in a travel planning task, a user (or group) can be recommended a package of places of interest (POI) that are within budget; i.e., which satisfy expressed constraints on budget or time [19, 20]. Travel recommender systems also need to be able to handle constraints, e.g. "no more than 3 museums" or "travel distance is less than 10 km" and provide alternatives for restaurants, transportation and hotels [1].

A second approach to package recommendations involves search. In the clothes domain, there are some package recommendation approaches based on image features [7, 13]. These approaches collect images (each image containing both a top and a bottom) from fashion websites [13] or fashion magazines [7] to create a package reference database. Using image processing techniques, they automatically separate the top and the bottom. Miura et al. [13] extracted image features (such as a RGB histogram and scale invariant features transform [SIFT] [10] values) for both tops and bottoms. To provide package recommendations, they required the user to provide a query (top or bottom) image. This image was then compared with packages in the reference database, and the closest package reference returned as a recommendation. Similar to Miura's work, Iwata et al. [7] extracted visual features (such as colour, texture and SIFT as a bag-of-features, and derived a topic model over these using Latent Dirichlet Allocation (LDA). When a user provided a query image (top/bottom), Iwata et al. recommended the other part by searching the topic model in their package reference database. Search can also take account of user context. Shen et al. [15] developed a clothes package recommendation system based on user context. First, they stored clothing items and combinations of items in a user wardrobe database. They also annotated its contents using

*Computing Science PhD Student at University of Aberdeen, UK. Lecturer at Telkom University, Indonesia, (agungtoto@telkomunivesity.ac.id)

User Modeling, Adaptation, and Personalization (UMAP) 2018, Singapore

DOI:

English words. To generate recommendations, their system asked the user about their goals ("destinations" and "want to look like") and mapped them to possible characteristics of clothes in the user wardrobe.

More recently, we suggested a collaborative filtering approach to package recommendations in the clothing domain using matrix factorization (MF) [18]. We showed that the user-package ratings matrix is too sparce to successfully apply MF methods. Instead, we applied matrix factorization separately to user-top and user-bottom rating matrices and predicted a package rating prediction by combining the predicted ratings for the top and bottom using either the minimum function or the harmonic means. This was the first collaborative approach to package recommendation, but had clear shortcomings in that it did not take into account fashion constraints that exist between the choice of top and bottom. For example, patterns and colors might clash between top and bottom, and they might not even be suitable for the same season. The work presented in [18] did not handle these scenarios.

In this paper, we propose a novel approach to package recommendation in the clothes domain that combines collaborative filtering on user-item matrices with constraints on the items within a package. To incorporate item constraints, we (a) enriched the dataset described in [18] by adding item attributes such as dress code, color and patterns through an annotation process; (b) constructed matrices of constraints (first hand-crafted and later through machine learning) on tops and bottoms within a package; and (c) proposed means to incorporate these constraints within matrix factorization to provide package recommendations.

The remainder of this paper is organized as follows. Section 2 defines the package recommendation task and the notation used, describes how the dataset was generated and enhanced, and formulates several matrix factorization approaches for package recommendation. Section 3 describes our motivations for using constraints, formulates how the handcrafted constraints were incorporated into matrix factorization for package recommendation, and describes how a supervised learning algorithm J48 [6] was used to learn constraints automatically. Section 4 details our experimental settings and Section 5 reports our experiment results. Section 6 describes the motivation, participants, materials, procedures, and findings of our focus group discussion to gain a deeper insight into how to improve cloths package recommendation. Finally, Section 7 provides a discussion and suggests directions for future work.

2 PACKAGE RECOMMENDATIONS

Consider a set of clothes $I^a = \{i^t_1, i^t_2, \ldots, i^t_o, i^b_1, i^b_2, \ldots, i^b_p\}$, consisting of two disjoint complementary sets: a set of o top items $I^t = \{i^t_1, i^t_2, \ldots, i^t_o\}$ and a set of p bottom items $I^b = \{i^b_1, i^b_2, \ldots, i^b_p\}$, where $I^t \cup I^b = I^a; o + p = n$.

Each item in I^a is associated with a set of q attributes $f^{1\cdots q}$, each of which takes one from a finite set of values, $\{f^1_{1\ldots r_1}, f^2_{1\ldots r_2}, ..., f^q_{1\ldots r_q}\}$.

Further, some of these items and their combinations (a package) have received ratings from one or more of m possible users $U = \{u_1, u_2, ..., u_m\}$. The individual ratings are denoted as a triple $(u, i, r_{u,i})$, where $u \in U$, $i \in I^a$ and $r_{u,i}$ is the rating given by user u to item i. Package ratings are denoted as a quadruple $(u, i^t, i^b, r_{u,(i^t,i^b)})$, where $u \in U$, $i^t \in I^t$, $i^b \in I^b$, and $r_{u,(i^t,i^b)}$ is the rating provided by user u to the package (i^t, i^b).

Our task is then to identify the best top-N package recommendations, based on both user–item and user–package ratings, and on the item features.

2.1 Dataset

In this paper, we extend the package recommendations dataset for the clothes domain [18]. This dataset contains 12,000 individual ratings and 6,000 package ratings from 200 users. The items consist of 1,400 "tops" and 600 "bottoms" extracted from Amazon product data [11, 12]. The dataset is publicly available and can be downloaded from a GitHub repository[1].

We further annotate all individual items by adding color, pattern and formality attributes. We use twelve different colors (black, gray, white, red, green, blue, yellow, orange, purple, brown, pink and other), seven different patterns (clean, text, checker, stripes, pattern, floral and picture) and four different formalities (casual, sport/outdoor, work and formal). Figure 1 shows examples of dress codes and patterns used in our dataset.

2.2 MF For Package Recommendations

Our starting point is previous work on MF for package recommendations [18]. There are two types of ratings matrices used in that study:

- CAT category ratings, where we use separate matrices for user–top ratings V^t and user–bottom ratings V^b
- ALL all ratings, where we use a single matrix for user–item ratings V^a.

There were four high performing solutions in [18], labeled MF-MIN-ALL, MF-MIN-CAT, MF-MUL-ALL, and MF-MUL-CAT. To understand these labels, each consists of three parts separated by a dash ("-") sign. The first part ("MF") describes that the solution is based on matrix factorization. The second part ("MIN" or "MUL") describes the use of minimum or multiplicative (harmonic mean) operations over individual rating predictions for top and bottom to generate a package rating prediction. The last part ("CAT" or "ALL") describes the use of separate matrices for top and bottom categories or a single matrix with all items, as described above.

The focus of this work is to combine these purely collaborative predictions with constraints pertaining to item features.

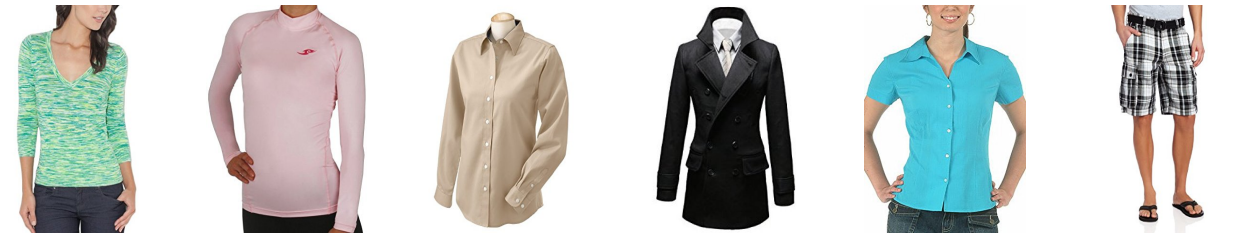
3 INCORPORATING CONSTRAINTS INTO MF

As discussed in Section 2.1, we enhanced our dataset by adding three attributes (color, pattern, and formality). In this section, we describe two methods (hand-crafted and automated) to determine the appropriateness of different combinations of tops and bottoms.

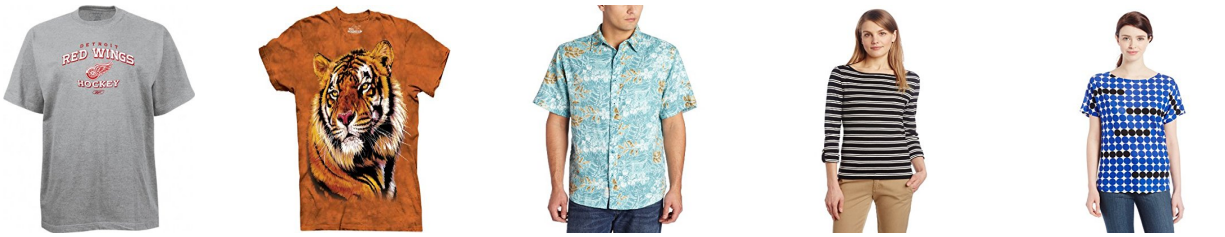
3.1 Hand Crafted Constraints

The annotation process as mentioned in Section 2.1 added color, pattern and formality attributes to each item in our dataset. Therefore, for each package rating we obtained information such as user (u), top item (i^t), bottom item (i^b), package rating $(u, i^t, i^b, r_{i^t,i^b})$, top

[1] https://github.com/atwRecsys/PackageRecDataset



(a) Casual (b) Sport/Outdoor (c) Work (d) Formal (e) Clean Pattern (f) Checker Pattern



(g) Text Pattern (h) Picture Pattern (i) Floral Pattern (j) Stripes Pattern (k) Pattern Pattern

Figure 1: Example of Images with Different Formality and Pattern Attributes

color (cl^{i^t}), top pattern (pt^{i^t}), top formality (fm^{i^t}), bottom color (cl^{i^b}), bottom pattern (pt^{i^b}) and bottom formality (fm^{i^b}).

3.1.1 Appropriateness Constraints Matrices. In order to identify the appropriateness of clothes combinations, we collected rules from fashion websites (i.e. Telegraph Fashion[2], AskMen[3], Looksgud[4], Effortless Gent[5], Gurl[6], Quora[7]). We collected statements pertaining to our attributes, and represented these as appropriateness matrices for each attribute. For example a statement of "*The best colors to wear together are shades that are complimentary of each other. ... These include red and green, violet and yellow and blue and orange.*" gives a clues that combination of red and green, or violet and yellow, or blue and orange works together.

For each attribute we experimented with three different matrices that we will refer to as: stick, carrot, and stick–carrot. We use the stick matrix to decrease the rating prediction when the top and bottom are identified as a non-appropriate combination, but if the combination is appropriate we do not give a reward. On the other hand, we use the carrot matrix to increase the rating prediction when the top and bottom are identified as an appropriate combination, but if the combination is not appropriate we do not give a penalty. When using the stick–carrot matrix, we apply either a reward or penalty to a combination depending on whether it is appropriate or not.

Table 1 shows example matrices for color, pattern and formality attributes. We provide an example each for carrot, stick and carrot–stick. A statement "White shirts go with everything", resulted in the values in the entire row for white tops having the value 1. Another statement "Keep red and pink separate" would have resulted in a value of -1 in the stick and carrot–stick matrices (not shown) for two cells: pink top and red bottom and vice versa.

3.1.2 Incorporating Color Constraint. To incorporate the color constraint into matrix factorization for package recommendations we follow Equation (1):

$$\hat{r}_{u_x,(i_y^t,i_z^b)} = f(u_x,(i_y^t,i_z^b)) + A^{cl}_{cl^{i^t},cl^{i^b}} * \rho^{cl} \quad (1)$$

where $f(u_x,(i_y^t,i_z^b))$ is the MF prediction for the package (using one of the algorithms MF-MUL-ALL, MF-MIN-ALL, MF-MIN-CAT, MF-MUL-CAT reported in [18]); $A^{cl}_{cl^{i^t},cl^{i^b}}$ is the prediction from the color appropriateness matrix for the top color (cl^{i^t}) and the bottom color (cl^{i^b}). Meanwhile, ρ^{cl} is the weight assigned to the color constraint prediction.

3.1.3 Pattern Constraint. Likewise, to incorporate pattern constraints into matrix factorization for package recommendations we follow Equation (2):

$$\hat{r}_{u_x,(i_y^t,i_z^b)} = f(u_x,(i_y^t,i_z^b)) + A^{pt}_{pt^{i^t},pt^{i^b}} * \rho^{pt} \quad (2)$$

where $A^{pt}_{pt^{i^t},pt^{i^b}}$ is the prediction from the pattern appropriateness matrix and ρ^{pt} is the weight assigned to the pattern constraint prediction.

3.1.4 Formality Constraint. Likewise, to incorporate the formality constraint into matrix factorization for package recommendations we follow Equation (3):

$$\hat{r}_{u_x,(i_y^t,i_z^b)} = f(u_x,(i_y^t,i_z^b)) + A^{fm}_{fm^{i^t},fm^{i^b}} * \rho^{fm} \quad (3)$$

[2]http://www.telegraph.co.uk/fashion/
[3]https://uk.askmen.com/style/
[4]https://www.looksgud.in/
[5]https://effortlessgent.com/
[6]http://www.gurl.com/
[7]https://www.quora.com/What-are-some-good-rules-of-thumb-for-women-when-putting-an-outfit-together

Table 1: Color appropriateness matrices (A), (a) Carrot appropriateness matrix for color, (b) Stick appropriateness matrix for pattern, and (c) Stick-Carrot appropriateness matrix for formality.

Carrot			cl^{i^b}											
			Black	Gray	White	Red	Green	Blue	Yellow	Orange	Purple	Brown	Pink	Others
														?
cl^{i^t}	Black		1	1	1	1	1	1	1	1	1	1	1	1
	Gray		1	1	1	1	1	0	1	1	0	0	1	0
	White		1	1	1	1	1	1	1	1	1	1	1	1
	Red		1	0	1	1	1	0	0	1	0	0	0	0
	Green		1	0	0	1	1	1	1	0	0	0	1	0
	Blue		1	0	0	0	1	1	0	1	1	0	0	0
	Yellow		1	0	1	0	1	0	1	0	1	0	0	0
	Orange		1	0	1	1	0	1	0	1	0	0	0	0
	Purple		1	0	0	1	0	1	1	0	1	0	0	0
	Brown		1	0	0	0	0	0	0	0	0	1	0	0
	Pink		1	0	1	0	1	0	0	0	0	0	1	0
	Others	?	1	0	0	0	0	0	0	0	0	0	0	1

(a)

Stick		pt^{i^b}						
		Clean	Text	Checker	Stripes	Pattern	Floral	Picture
pt^{i^t}	Clean	0	0	0	0	0	0	0
	Text	0	-1	-1	0	-1	-1	-1
	Checker	0	-1	-1	0	-1	-1	-1
	Stripes	0	0	0	0	0	0	0
	Pattern	0	-1	-1	0	-1	-1	-1
	Floral	0	-1	-1	0	-1	-1	-1
	Picture	0	-1	-1	0	-1	-1	-1

(b)

Stick-Carrot		fm^{i^b}			
		Casual	Sport/Outdoor	Work	Formal
fm^{i^t}	Casual	1	1	1	-1
	Sport/Outdoor	1	1	-1	-1
	Work	-1	-1	1	1
	Formal	-1	-1	1	1

(c)

where $A^{fm}_{fm^{i^t}, fm^{i^b}}$ is the prediction from the formality matrix and ρ^{fm} is weight assigned to the formality constraint prediction.

3.1.5 Incorporating Multiple Constraints. In Equation (1), (2) and (3), the color, pattern and formality constraints are only incorporated individually. We also experiment with combining two or three constraints together by using Equation (4).

$$\hat{r}_{u_x,(i^t_y,i^b_z)} = f(u_x,(i^t_y,i^b_z)) + A^{cl}_{cl^{i^t},cl^{i^b}} * \rho^{cl} + A^{pt}_{pt^{i^t},pt^{i^b}} * \rho^{pt} + A^{fm}_{fm^{i^t},fm^{i^b}} * \rho^{fm} \quad (4)$$

For instance, the combination of color and pattern is achieved by setting the formality appropriateness wight (ρ^{fm}) to 0.

3.2 Automatically Learned Constraints

As an alternative to hand-crafting attribute constraints from fashion literature, we also explored the option of learning constraints from the dataset.

In this scenario, we consider each package rating $(u, i^t, i^b, r_{i^t,i^b})$ in the training set, and:

(1) discard the user u;
(2) represent the top and bottom by their attributes (top color cl^{i^t}, pattern pt^{i^t} and formality fm^{i^t}; bottom color cl^{i^b}, pattern pt^{i^b} and formality fm^{i^b});
(3) convert the package rating r_{i^t,i^b} into a binary label "good" (rating of 4–5) or "bad" (rating of 1–3); and

Then we train the J48 classifier [6] over this converted training set to classify packages as good (1) or bad (1). To obtain a rating prediction, we take the strength of the prediction of the classifier $pred(i^t_y, i^b_z)$, whose value ranges from 0 (bad) to 1 (good), and normalize this by the following Equation 5.

$$g_{(i^t_y,i^b_z)} = min + (max - min) \times pred(i^t_y, i^b_z) \quad (5)$$

where min and max are the minimum and maximum rating values (in our dataset, 1 and 5).

3.2.1 Incorporating Decision Trees with MF. To get ranking predictions, we combine the MF-MUL-CAT prediction with the J48 prediction as follows:

$$\hat{r}_{u_x,(i^t_y,i^b_z)} = f(u_x,(i^t_y,i^b_z)) * pred(i^t_y,i^b_z) \quad (6)$$

where the $f(u_x,(i^t_y,i^b_z))$ is the rating prediction produced by the matrix factorization adaptation, and the $pred(i^t_y,i^b_z)$ is the strength of the predictions produced by the decision tree/J48 classifier (see Equation (5)).

4 EXPERIMENTAL SETTINGS

4.1 Evaluation Metric

In this paper we evaluate based on a top-N package recommendation scenario, where we evaluate the quality of the top N recommendations we make for different users. This is a more realistic evaluation that those based on average rating prediction accuracy using metrics, e.g. root mean squared error (RMSE), which give undue importance to items that are not being recommended. There are several metrics that can be used to determine top-N performance, e.g. Normalised Discounted Cumulative Gain (NDCG), recall@N, precission@N, etc. In this paper, we adopt recall@N as described by Cremenosi et.al [3], as its assumptions are a good fit for package recommendations. We will say more about this after describing the method.

The recall@N metric is calculated as follows: First, from the testing set we collect all the packages (i^t, i^b) rated 5 by each user u into a set T. Second, for each package contained in T:

(1) We select 99 random packages and assume that the user u will not like these packages as much;

(2) We predict the ratings given by user u for the test package (i^t, i^b), which has a known score of 5, and for the 99 additional packages, which are assumed to be rated lower;
(3) We form a ranked list by ordering all 100 packages according to the rating predictions. Let p denotes the rank of the test package (i^t, i^b) within the list.
(4) We form a top-N recommendation list by picking the N top ranked packages from the list. If $p \leq N$ we have a *hit*, otherwise we have a *miss*. Chances of a hit increase with an increase in the N value and are guaranteed at $N = 100$.

Average Recall@N performance is then defined by Equation 7:

$$recall(N) = \frac{\#hit}{|T|} \tag{7}$$

4.2 Crossvalidation Method

We apply a 4-fold crossvalidation methodology. This is done by randomly splitting the individual ratings into four parts, and then rotating and using three parts as the training set and one for testing. Following [18], in each fold we used only 25% of package ratings $r_{u,(i^t,i^b)}$ as the training set, and the remaining 75% package ratings $r_{u,(i^t,i^b)}$ as the test set.

4.3 Experimental Settings

Wibowo et.al [18] cast package recommendation as a rating predictions task. In their work they used RMSE as a performance metric. Since in this paper we use the top-N package recommendation as scenario, to get a fair comparison we first reran those algorithms (MF-MIN-ALL, MF-MIN-CAT, MF-MUL-ALL, and MF-MUL-CAT) and report the top-N package performance. In this paper, we will use these algorithms as baselines.

To obtain a better understanding on how our adaptations produce improvements, we report the recall@10 performance over a combination of each constraint with the best baseline performance from above:

(1) We report a combination of the best solution (from our baselines) with each constraint attributes (color, pattern and formality) as described in Section 3.1.2. We report results for the best performing value of the weights ρ=0.8.
(2) We report combinations of constraints which produced an improvement using the scenario as explained in Section 3.1.5.
(3) We report using the decision tree algorithm (Section 3.2).
(4) We report the combination of the decision tree algorithm with the best performing MF baseline.

5 RESULTS

5.1 MF For Package Recommendation Baseline

Table 2 shows the average recall@10 on the testing set for different algorithms described in [18]. The green cell in this table shows the best recall@10 from purely collaborative approaches. In this section we will use this algorithm MF-MUL-CAT as our baseline and incorporate constraints as described in Section 3.1 into MF-MUL-CAT for comparison.

Table 2: Average Recall@10 Performance of the baseline.

Scenario	recall@10
MF-MIN-ALL	0.1220
MF-MIN-CAT	0.1147
MF-MUL-ALL	0.1253
MF-MUL-CAT	0.1293

5.2 Incorporating Handcrafted Constraints Individually

Table 3 shows the average recall@10 for combining MF-MUL-CAT with handcrafted rules. The column "Attribute" denotes one of our attributes (color, pattern, and formality) as mentioned in Section 2.1. The column "Type" denotes one of our constraint matrix types (Carrot, Stick, and Stick–Carrot) in each attribute as mentioned in Section 3.1.

Table 3: Average Recall@10 Performance for MF with Handcrafted Constraints

Scenario	Attribute	Type	recall@10
MF-MUL-CAT			0.1293
MF-MUL-CAT	Color	Carrot	0.1296
MF-MUL-CAT	Color	Stick	0.1391
MF-MUL-CAT	Color	Stick–Carrot	0.1332
MF-MUL-CAT	Pattern	Carrot	0.1260
MF-MUL-CAT	Pattern	Stick	0.1364
MF-MUL-CAT	Pattern	Stick–Carrot	0.1245
MF-MUL-CAT	Formality	Carrot	0.1532
MF-MUL-CAT	Formality	Stick	0.1461
MF-MUL-CAT	Formality	Stick–Carrot	0.1522

As we can see from Table 3, all but two of our adaptations outperform the MT-MUL-CAT baseline (the yellow cell in the recall@10 column). The green cells in Table 3 represent the best combinations for each attribute, and we can see that incorporating constraints for each of our attribute types improves performance compared to a purely collaborative approach, with constraints on formality providing the biggest gain.

5.3 Incorporating Combinations of Handcrafted Constraints

Table 4 shows the average recall@10 for different combinations of handcrafted rules. The first four rows in Table 4 summarise our best recall@10 for each attribute in Table 3. The remaining four rows report combinations of those attributes following the formula in Section 3.1.5.

As we can see from Table 4, all of our adaptations outperform the MF-MUL-CAT baseline. The best performance comes from incorporating all three attributes (the green cell with a recall@10 value of 0.1604).

5.4 Decision Tree Performance

Table 5 shows the average recall@10 using the decision tree (J48). The column "Upsampling" shows the factor used for upsampling

Table 4: Average Recall@10 Performance for MF with combinations of handcrafted constraints

Scenario	Attribute	App. Type	recall@10
MF-MUL-CAT			0.1293
MF-MUL-CAT	Color	Stick	0.1391
MF-MUL-CAT	Pattern	Stick	0.1364
MF-MUL-CAT	Formality	Carrot	0.1532
MF-MUL-CAT	Comb. of Color and Pattern		0.1398
MF-MUL-CAT	Comb. of Color and Formality		0.1548
MF-MUL-CAT	Comb. of Pattern and Formality		0.1488
MF-MUL-CAT	Comb. of All Attributes		0.1604

the minority class of good packages (defined as those for which $r_{u,(i^t,i^b)} = 4, 5$) in our traning set. The first two rows show the recall@10 for MF-MUL-CAT and also the best combination of MF-MUL-CAT with handcrafted constraints, summarized from previous tables. The next three rows report results using only the learned constraints, i.e. with no collaborative element. As expected these are worse than the results of collaborative filtering, highlighting that user preferences play a key role in this domain. The last three rows report results for the combination of J48 predictions with the baseline MF approach. While there is an improvement over the baseline, the automatically learned constraints do not perform as well as the manually curated ones.

Table 5: Average Recall@N Performance using Automatically Learned Constraints

Scenario	Upsampling	recall@10
MF-MUL-CAT		0.1293
MF-MUL-CAT (All Attributes)		0.1604
J48	2	0.0990
J48	3	0.1016
J48	4	0.0981
J48 * MF-MUL-CAT	2	0.1439
J48 * MF-MUL-CAT	3	0.1509
J48 * MF-MUL-CAT	4	0.1477

To summarize, we have reported several results for package recommendations by combining matrix factorization predictions with constraints about which attributes can go well together and which ones can clash. To better understand our results and also gain insights as to how to improve the system for the future, we conducted focus groups on package recommendations, described next.

6 FOCUS GROUPS ON CLOTHES COMBINATION ASPECTS

The aim of our focus groups (FGs) was to gain a better understanding of what aspects affect whether clothes are good to combine together. We were interested in people's arguments and preferences for combining a "top" (e.g. a shirt or t-shirt) and "bottom" (e.g. trousers, shorts, skirts). We were also interested in the clothing features that result in a positive, negative or neutral judgment on a particular clothing combination. This study leads to a deeper insight into how to improve package recommendation in the clothes domain, such as what constraints may be better to use and how to use them.

6.1 Participants

Eight FGs were held with 3-5 participants per group. 30 participants were recruited (14 female and 16 male) using convenience sampling from the University of Aberdeen. They came from 11 different countries. All were over 18; further demographic data was not collected.

6.2 Materials

We ran our FGs in two scenarios. In the first scenario, we showed the FGs some images of "top" and "bottom" clothes and their ratings by an anonymous user. We selected the clothes at random, and varied the number of clothes shown to the participants (see Table 6). We also showed 3 combinations of tops and bottoms to the participants. In the second scenario, we showed the FGs 15 combinations, selected at random.

For our anonymous users, we randomly selected 4 male and 4 female users from our previous dataset[8]. To make the discussion easier, we provided pseudonames in each scenario: Alice, Barbara, Carol, and Deborah as female pseudonames, and Andy, Bob, Charles, and David as male pseudonames.

Table 6: Frequency Distribution of Individual Preferences Samples Involved in FGD

Pseudoname	Gender	#Tops at Rating					#Bottoms at Rating				
		1	2	3	4	5	1	2	3	4	5
Alice	Female	2	2	2	2	4	3	1	0	4	4
Andy	Male	2	2	1	2	3	0	0	1	2	4
Barbara	Female	0	1	3	4	3	1	4	1	1	2
Bob	Male	1	3	3	4	2	1	3	2	2	4
Carol	Female	1	1	3	3	3	3	3	2	3	3
Charles	Male	4	1	2	2	2	2	1	2	1	1
Deborah	Female	3	2	3	3	4	2	2	2	2	1
David	Male	3	0	3	3	3	3	1	2	1	3

6.3 Procedure

Participants were told the purpose of the FG was to understand what aspects affect whether clothes are good to combine together. Next, the FGs were run using the following steps:

Step 1. The FG was given the user preferences, e.g with pseudoname is "Alice", and asked to discuss whether "Alice" will like/dislike each combination from the 3 given pairs. The FG was also asked to identify why "Alice" will like/dislike the particular combination.

Step 2. With the first scenario still visible, 15 combinations were provided for "Alice" and the FG was asked to select the 3 combinations that "Alice" might like the best. We also asked them to identify what makes these the best combinations for "Alice".

Step 3. Participants were invited to share any other thoughts they had related to clothes combination aspects.

[8]https://github.com/atwRecsys/PackageRecDataset

Step 4. Step 1-3 were repeated for another user with different gender.

FGs were audio recorded, and a thematic analysis was conducted based on these recordings.

6.4 Results

Even though we conducted our discussion using two scenarios, we do not distinguish our findings into two separate tables because both scenarios are intended to provide insights into what aspects affect whether clothes are good to combine together.

Table 7 summarizes the finding on clothes combination aspects. This table contains 9 columns, with the first column indicating aspects mentioned by the FGs. The other columns represent the FGs. Each FG is named by taking the first letter from the pseudonames discussed in it and adding the FG sequence number. Therefore, column "C5" represents FG number 5 which discussed users "Carol" and "Charles". The user pseudonames are listed in Table 6. The check-mark (✓) indicates that a particular aspect was mentioned in the FG discussion.

Table 7: Clothes Combination Aspects Discussed in the Study

Aspects		A1	A2	B3	B4	C5	C6	D7	D8
Combination Aspects									
Formality agreement		✓	✓	✓	✓	✓	✓	✓	✓
Use the high rating		✓	✓	✓	✓	✓	✓	✓	✓
Color compatibility		✓	✓		✓	✓	✓	✓	✓
Brightness composition			✓	✓	✓	✓		✓	✓
Eliminate the low rating		✓	✓	✓			✓		✓
Motif harmony					✓	✓		✓	✓
Length of the dress			✓		✓				
Functional agreement			✓		✓				
Seasonal agreement								✓	
Individual Preferences									
Color	Favorite color	✓	✓	✓	✓	✓	✓	✓	✓
	Color brightness		✓	✓	✓	✓	✓	✓	✓
Style	Sleeve length	✓	✓	✓	✓	✓	✓	✓	✓
	Bottom length	✓	✓	✓	✓	✓	✓	✓	✓
	Cutting shapes	✓	✓	✓	✓	✓	✓	✓	✓
	Shirt type	✓	✓	✓	✓	✓	✓		
	Loose style			✓	✓	✓		✓	✓
	Collar			✓	✓		✓		✓
	Body exposure			✓		✓	✓		
	Formality		✓		✓	✓			
Patterns		✓	✓		✓	✓	✓	✓	✓
User personality		✓			✓	✓			✓
Fabric			✓		✓	✓			
Cloth details					✓			✓	

In Table 7, we group the aspects findings into two different sets: combination aspects and individual preferences. Even though there are similar aspects that appear in both sets, participants tended to use them in different situations. For example, in statement "*agree on combinations six [..] six is the formal shirt and the formal trousers*" the formality aspect is more about the combination rather than individual preferences. In contrast, a statement such as "*He doesn't quite like the more formal shirt [..]. So in all I wouldn't say yes or no to this shirt*" is more about individual preferences.

6.4.1 Combination Aspects. From the FGs, we identified some combination aspects that affect whether clothes are good to combine together. As mentioned in Table 7, there are 9 of these:

(1) Formality agreement (all FGs). Participants conveyed that a clothes combination must satisfy the dress code formality agreement. They also identified formal, semi-formal/work, casual, and sport as types of formality.
(2) Use high ratings (all FGs). Participants argued that people tend to use clothes from the items they love when creating a clothes combination. The set of loved items can also be expanded to include items of similar type and color.
(3) Color compatibility (7 FGs). Participants argued that a good combination should have good color compatibility. A combination from gradation colors was considered as a good combination. They also mentioned some conflicts caused by colors when clothes are combined together.
(4) Brightness composition (6 FGs). Most participants who considered the brightness composition aspect said that the composition of bright and dark or bright and neutral for clothes works well together.
(5) Eliminate low ratings (5 FGs). This aspect was strongly expressed when the FGs were asked to find 3 out of 15 as the best combinations. They easily discarded combinations which contained lowly rated items. They felt that the combination was ruined when it contained any disliked item.
(6) Motif harmony (4 FGs). Some participants expressed that it will be better for a combination to have patterns only in one part of clothing. The pattern can be in the top or the bottom as long as the other part is plain.
(7) Length of dress (2 FGs). Participants noted that there are some dresses that do not need a bottom.
(8) Functional agreement (2 FGs). Some participants argued that we can classify clothes into some functional categories e.g go to beach. Some clothes combinations are better when they are in the same category.
(9) Seasonal agreement (1 FG). Participants argued that a winter jacket and summer pants did not work together.

6.4.2 Individual Preferences. The individual preferences described the reason why a user like/dislike a particular item. These individual preferences might affected user judgment to the whole combinations. As shown in Table 7, there are 14 matters involved in individual preferences, which can be grouped into categories.

Color.

- *Favorite color* (all FGs). Participants argued that some users tend to select clothes based on their favorite color. These colors were identified from the samples provided with ratings 4 and 5
- *Color brightness* (7 FGs). Participants argued that some users tend to select clothes with similar brightness (e.g.

neutral, calm, dark, bright). Here, we distinguish the favorite color from the color brightness since in the discussion there were participants who believed that the user did not have any objection to bright blue and dark blue.

Style. Participants discussed many different aspects of style, far more than just the distinction between formal and informal that we had been making.

- Sleeve length (all FGs). Respondents argued that some users tends to choose clothes with a particular length of sleeve.
- Bottom length (all FGs). Participants argued that some users tends to choose a bottom with a particular length (e.g. capri jeans, sort jeans).
- Cutting shapes (all FGs). Participants argued that some users tend to choose bottoms with a particular cutting shape e.g. bell bottom, boot cut, slim fit (tight), baggy pants.
- Shirt type (6 FGs). Participants argued that some users have preferences for the shirt type (e.g. polo, T-shirt, long T-shirt, shirt, formal shirt, hoodie).
- Loose style (5 FGs). Participants described the loose style as clothes which are extra wide at the bottom, and they believed that these aspects influence user preferences. They added that a loose cloth style provides freedom to a person wearing it.
- Collar (4 FGs). Participants argued that some users tend to wear/avoid clothes with or without collar.
- Formality (3 FGs). Participants argued that some users tends to choose individual items based on formality. Even though formality agreement was discussed by all FGs, the formality aspect in individual preferences was not discussed by all. Participants argued that some users would like to select different formality to wear to different occasions.
- Body exposure (3 FGs). Participants argued that some users prefer to wear/avoid clothes which expose some part of their body.

Patterns (7 FGs). Participants argued that some users love to wear clothes with patterns (e.g. checker) either on top or bottom.

User personality (5 FGs). Participants argued that a user's personality (e.g. mature, easy going) affects the selection of individual items. A mature user prefers to select a particular type of clothes.

Fabric (3 FGs). Participants argued that some users did not have any objection to clothes with thick fabric (e.g jeans), or thin/loose fabric (e.g cotton, silk, or satin).

Cloth details (2 FGs). Participants argued that some users prefer to wear/avoid clothes with tiny details on them (e.g. gold buttons on a black shirt).

7 DISCUSSION AND FUTURE WORK

We have present a novel approach to package recommendations by incorporating constraints on the the item attributes into a matrix factorization based collaborative filtering algorithm. Our results show that modeling constraints, either through manual curation from external resources, or through automated acquisition from within the dataset, is an important step in the cloths domain.

To gain further insights into package recommendation in this domain, we conducted focus groups that revealed two types of considerations: combination aspects and individual preferences. The combination aspects reflect the reasons why a user might like or dislike a combination of clothes, while the individual preferences identify item features that are relevant for modeling whether a user will like or dislike a top or a bottom individually. The focus group validated several aspects of our algorithm, for instance, the importance of considering colors, formalities, and patterns, our choice of minimum and harmonic mean operations for combining individual ratings for tops and bottoms. None the less, there ware aspects which not included in our adaptations, such as brightness composition, length of the dress, functional agreement and seasonal agreement.

The focus groups also identified several features that affect user preferences, for example, favorite colors, brightness, lenghts, cuts and looseness of outfits, formality, fabric. We do not explicitly model these features in our algorithm and instead rely on matrix factorization to identify latent item and user features from the training data.

Our work can immediately be extended in a couple of ways. One is take into account our focus group findings which consider the individual preferences and user attributes explicitly. The second is to expand the number of item features modeled for purpose of constraining potential combinations of clothes.

In future work we also propose to extend our model to handle other types of constraints, for instance, budget, and to allow for packages with larger number of items. In this context, we would also like to investigate the package recommendation challenge in other domains, for example food or travel.

ACKNOWLEDGMENTS

We would like to thank Lembaga Pengelola Dana Pendidikan (LPDP), Departemen Keuangan Indonesia for awarding a scholarship to support the studies of the lead author. We would also like to thank the participants in our focus groups who communicated precious feedback during our discussions.

REFERENCES

[1] Sihem Amer-Yahia, Francesco Bonchi, Carlos Castillo, Esteban Feuerstein, Isabel Mendez-Diaz, and Paula Zabala. 2014. Composite retrieval of diverse and complementary bundles. *IEEE Transactions on Knowledge and Data Engineering* 26, 11 (2014), 2662–2675.

[2] Da Cao, Liqiang Nie, Xiangnan He, Xiaochi Wei, Shunzhi Zhu, and Tat-Seng Chua. 2017. Embedding Factorization Models for Jointly Recommending Items and User Generated Lists. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 585–594.

[3] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 39–46.

[4] Michael D Ekstrand, John T Riedl, and Joseph A Konstan. 2011. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction* 4, 2 (2011), 81–173.

[5] A Felfernig, S Gordea, D Jannach, E Teppan, and M Zanker. 2007. A short survey of recommendation technologies in travel and tourism. *OEGAI journal* 25, 7 (2007), 17–22.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

[7] Tomoharu Iwata, Shinji Wanatabe, and Hiroshi Sawada. 2011. Fashion coordinates recommender system using photographs from fashion magazines. In *IJCAI*, Vol. 22. Citeseer, 2262.

[8] Qi Liu, Enhong Chen, Hui Xiong, Yong Ge, Zhongmou Li, and Xiang Wu. 2014. A cocktail approach for travel package recommendation. *IEEE Transactions on Knowledge and Data Engineering* 26, 2 (2014), 278–293.

[9] Pasquale Lops, Marco Gemmis, and Giovanni Semeraro. 2011. Recommender Systems Handbook. Content-based Recommender Systems: State of the Art and Trends (2011), 73–105. http://dx.doi.org/10.1007/978-0-387-85820-3_3

[10] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, Vol. 2. Ieee, 1150–1157.

[11] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

[12] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.

[13] Shinya Miura, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2013. SNAPPER: fashion coordinate image retrieval system. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on*. IEEE, 784–789.

[14] X. Ning, C. Desrosiers, and G. Karypis. 2015. *A comprehensive survey of neighborhood-based recommendation methods*. 37–76.

[15] Edward Shen, Henry Lieberman, and Francis Lam. 2007. What am I gonna wear?: scenario-oriented recommendation. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 365–368.

[16] Xiaoyuan Su and Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009 (2009), 4.

[17] A. T. Wibowo, A. Siddharthan, H. Anderson, A. Robinson, Nirwan Sharma, H. Bostock, A. Salisbury, R. Comont, and R. V. D. Wal. 2017. Bumblebee Friendly Planting Recommendations with Citizen Science Data. In *Proceedings of the RecSys 2017 Workshop on Recommender Systems for Citizens co-located with 11th ACM Conference on Recommender Systems (RecSys 2017), Como, Italy, August 31, 2017*.

[18] A. T. Wibowo, A. Siddharthan, C. Lin, and J. Masthoff. 2017. Matrix Factorization for Package Recommendations. In *Proceedings of the RecSys 2017 Workshop on Recommendation in Complex Scenarios co-located with 11th ACM Conference on Recommender Systems (RecSys 2017), Como, Italy, August 31, 2017*. 23–28. http://ceur-ws.org/Vol-1892/paper5.pdf

[19] Min Xie, Laks VS Lakshmanan, and Peter T Wood. 2010. Breaking out of the box of recommendations: from items to packages. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 151–158.

[20] Min Xie, Laks VS Lakshmanan, and Peter T Wood. 2011. Comprec-trip: A composite recommendation system for travel planning. In *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*. IEEE, 1352–1355.